Individual Diploma Thesis

ARTIFICIAL INTELLIGENCE AND ETHICS

NICOLAS IOANNOU

UNIVERSITY OF CYPRUS

DEPARTMENT OF COMPUTER SCIENCE

May 2022

UNIVERSITY OF CYPRUS

DEPARTMENT OF COMPUTER SCIENCE

Artificial Intelligence and Ethics

Nicolas Ioannou

Supervisor

Prof. Keravnou-Papailiou Elpida

The Individual Diploma Thesis was submitted for partial fulfilment of the requirements for obtaining the degree of Computer Science of the Department of Computer Science of the University of Cyprus

May 2022

Acknowledgments

I would like to thank my thesis supervisor Prof. Keravnou-Papailiou Elpida, for the valuable guidance, encouragement, and support during this project. Her contribution and feedback were critical to complete and write this dissertation.

I would also like to thank my family and friends for all the support they have provided throughout the years of my education.

Περίληψη

Η Τεχνητή Νοημοσύνη έχει εισχωρήσει σε κάθε κομμάτι της ζωής μας την τελευταία δεκαετία. Από τον έξυπνο βοηθό στα smartphone μας μέχρι και στην εκπαίδευση, την αστυνόμευση και ακόμη και την ιατρική. Το γεγονός ότι η τεχνητή νοημοσύνη έφερε επανάσταση σε αυτούς τους τομείς είναι αδιαμφισβήτητο και δεν φαίνεται να επιβραδύνεται. Γι' αυτό είναι σημαντικό να κάνουμε ένα βήμα πίσω και να κατανοήσουμε ότι η τεχνητή νοημοσύνη δεν είναι η πανάκεια για όλα τα προβλήματά μας. Αλλά μάλλον ένα πρόβλημα από μόνο του.

Ο στόχος της εργασίας μου είναι να διερευνήσω μερικές από αυτές τις τελευταίες εξελίξεις στην Τεχνητή Νοημοσύνη που θέτουν ηθικούς προβληματισμούς. Δηλαδή, αυτόνομα οχήματα, Google Duplex και ευφυείς διεπαφές χρήστη. Εκτός από την αναγνώριση των πολυάριθμων ηθικών ανησυχιών, ήθελα επίσης να παρουσιάσω και να επικρίνω διαφορετικές λύσεις που προτάθηκαν από ειδικούς.

Αφού αναλύσουμε κάθε πρόοδο και συζητήσουμε τις πολιτικές και τις κατευθυντήριες γραμμές που έχουν καθιερωθεί προηγουμένως, διερευνούμε πώς το Responsible AI είναι η καλύτερη προσέγγιση για την επίλυση πολλών από τα ηθικά ζητήματα που αναφέρθηκαν σε όλη τη εργασία μου.

Στο τέλος κάνω μερικές προτάσεις για το πώς να προσεγγίσουμε αυτά τα θέματα και πώς να διασφαλίσουμε την ηθική τεχνητή νοημοσύνη.

Abstract

Artificial Intelligence has weaved its way into every part of our lives in the last decade. From the smart assistant on our smartphones to having real life impact on education, policing and even healthcare. The fact that AI has revolutionized these sectors is irrefutable, and it does not seem to be slowing down. That is why it is important to take a step back and understand that AI is not the panacea to all our problems. But rather a problem in it of itself.

The goal of my thesis is to explore some of these latest advances in Artificial Intelligence that pose ethical concerns. Namely, Autonomous Vehicles, Google Duplex, and Intelligent User Interfaces. Apart from acknowledging the numerous ethical concerns, I wanted to also showcase and criticize different solutions that were suggested by experts.

After taking a deep dive on each advancement and discussing previously established policies and guidelines, we explore how Responsible AI is the best approach to solve many of the ethical issues that were brought up throughout the thesis.

In the end I make some suggestions on how to approach these subjects and how to ensure ethical AI.

Contents

Chapter 1

Introduction

1.1 Overview

Artificial Intelligence has weaved its way into every part of our lives in the last decade. From the smart assistant on our smartphones to having real life impact on education, policing and even healthcare. The fact that AI has revolutionized these sectors is irrefutable, and it does not seem to be slowing down. That is why it is important to take a step back and understand that AI is not the panacea to all our problems. But rather a problem in it of itself.

A lot of the latest AI developments have had significant backlash from the general public and the scientific community. To be more exact, many ethical concerns were raised about the use of AI in some aspects of our daily lives, such as autonomous vehicles. The concerns that were raised are very valid and act as a wake-up call for AI developers and engineers.

AI has made our life easier and more accessible. We want to keep benefiting from it whilst also making sure that we use it ethically. Not doing so could end up catastrophic not only for future AI developments, but also humankind.

1.2 Purpose

The goal of my thesis is to explore some of these latest advances in Artificial Intelligence that pose ethical concerns. I chose to mention these advances because they sparked my interest when they were announced, and I found the discourse surrounding them fascinating.

Throughout my thesis I express the public's opinion on these matters, but also shine a light on the opinion of experts and how these different opinions helped me shape mine. Apart from acknowledging the numerous ethical concerns, I wanted to also showcase and criticize different solutions that were suggested by experts.

1.3 Outline

The rest of my thesis is organized in chapters as follows.

In Chapter 2, I provide the necessary information to understand what Artificial Intelligence and Ethics are. Here the three main ethical theories are explained, and answers are provided on if morality is innate or learned and how we can code it.

In Chapter 3, the concept of autonomous vehicles is explored. After giving a definition for them, but also explaining what the 'trolley problem' is, we see one of the first solutions that was suggested, 'The Moral Machine'. After expounding on why it was a wrong approach we give the spotlight to a more correct approach, Germany's Act on Autonomous Driving.

In Chapter 4, we take a deep dive in two Google products, Duplex and Screen Call. We see the backlash these advances have received and discuss two different approaches on how we should view Duplex. In the end we see how Google has responded and provide ethical guidelines for future advances.

In Chapter 5, we talk about the impact of Intelligent User Interfaces, their ethical concerns and what guidelines are in place.

In Chapter 6, we explore how Responsible AI is the answer to many of these ethical concerns and take a deep dive on Europe's 'Ethical Guidelines for Trustworthy AI'.

Finally, in Chapter 7, I summarize what was said and discuss the outcomes of my thesis.

Chapter 2

Background

2.1 Artificial Intelligence

The definition of Artificial Intelligence, just like many things in life, has changed a lot throughout time [1]. The European Commission defines AI as "a system that displays intelligent behavior by analyzing their environment and taking actions to achieve specific goals" [2]. Some examples of such systems are voice assistants, search engines and autonomous cars. Another definition we could use is that AI is an artificial agent that thinks or acts like humans [3]. This definition was used in the founding document that established the field of AI, 'Proposal for the Dartmouth Summer Research Project on Artificial Intelligence', by McCarthy et al.

We can divide AI in two categories, Weak AI, and Strong AI. Their main difference is that Weak AI is limited to a single narrowly defined task, whereas Strong AI can be occupied with multiple tasks which are not narrowly defined. [1]

In 1950 Alan Turing suggested a way to test whether a machine is considered intelligent or not. The now famous Turing Test states that a machine can be regarded as intelligent if after using written communication, the person interacting with it cannot determine whether the machine is a person or not. [1, 3]

No machine has passed the Turing Test to this day and no AI has achieved Strong AI capabilities. But some systems have shown Weak AI capabilities like beating an experienced player in Go or chess [1]. Unfortunately, even if a system passes the Turing Test, nobody can confirm that the passing is a required or adequate condition for intelligence [1], as intelligence is a vague concept [2].

2.1.1 The Philosophy View

Questions like “What does it mean for a machine to act intelligently?” and “What are the differences between machine intelligence and ours?” have long been the focus of philosophical reflection. Computer scientists are concerned with how to build and develop concepts like intelligence and consciousness, whereas philosophers ask what these concepts actually mean. Philosophy also contemplates on the question of whether intelligence requires consciousness, which most philosophers believe that it does not. Finally, philosophers like Nick Bostrom, reflect on the concept of singularity, meaning the point where AI will trigger uncontrolled technological growth that might even lead to the human extinction. [3]

2.2 Ethics

When thinking of ethics, the first thing that comes to our mind is if something is ‘morally good’ or ‘morally bad’ or to put it simply, ‘right’ or ‘wrong’ [1]. Generally speaking, ethics is concerned with principles and norms [1], which are rules that interpret “you should” kind of statements [4]. Ethics are believed to have originated

from Greece over two thousand years ago from Socrates. Over the years, philosophers have developed many ethical theories, which are theories on how to act and take decisions [5]. We will focus on three of them: Deontological, Consequentialist and Virtue ethics.

2.2.1 Deontological Ethics

The term deontological derives from the Greek word 'δέον', which means obligation or duty [1]. This is what this theory is based on. The ethical correctness of an action is evaluated by the intention of said action [1] or whether it is in accordance with a set of principles, like 'do not lie' etcetera [5]. A famous deontologist, Immanuel Kant, believed that for a moral evaluation of our actions, consequences should not matter [4]. We cannot control the future, so we should be lauded or condemned for actions within our control and not our achievements [4].

2.2.2 Consequentialist Ethics

In contrast to deontology, in consequentialism the ethical correctness of an action is evaluated solely based on its consequences [1]. There three main types of consequentialism theory: Utilitarianism, Egoism and Altruism [6]. In Utilitarianism an action is considered good if it benefits the maximum number of people whereas in Egoism an action is good so long as it maximizes an individual's happiness [6]. Lastly in Altruism, an action is good if it benefits everyone except the actor [6].

2.2.3 Virtue Ethics

In Virtue Ethics an action is considered good if it were what a virtuous agent would do in the situation at hand, where a virtuous agent is someone who acts virtuously. Virtue in Greek means excellence of a person. So, this theory actually has to do with building good character, and according to Plato, a person can do that by acting with wisdom, courage, temperance, and justice. If one has and acts with those traits, then good actions will follow. [4]

2.2.4 Is it innate and can it be learned?

Why act in a moral way? Why be good? The answers vary based on who you ask. Deontologists will say that it is our duty to act good, virtue ethicists that it is for improving our moral character, while religious people will say that it is what God commanded. [7]

But what if we lived in a world where acting immorally would have zero consequences and our actions were only based on our self-interest? In this imaginary world, ethics would lose its raison d'etre and the answer to our previous questions would be that there is no reason to act good unless we have self-interest. [7]

So, let us assume that we do have reason to be good and act morally. Did we learn to be moral, or have we been born with a sense of morality? Though we are not born good, children do appear to show empathy and altruism as shown by their will to share or help a stranger [7]. This is according to Paul Bloom, who is the person behind this research. He also argued that since children are too young to have learned about morality, that means that we have an innate sense of morality. This conclusion though raises another question.

Is morality universal? Certain values do appear to go beyond culture and location, such as ‘helping someone in need is good’ and ‘cheating is bad’ [7]. We will see in a later section that morality is not as universal as we would want it to be.

Socrates, a Greek philosopher, debated whether ethics can be taught. He believed that ethical knowledge can be taught, and that ethical behavior can be developed [8]. Kant also believed that ethics were a learned behavior that could evolve as we grow [9].

To simplify it, our ethical judgment develops throughout life, through our experiences and interactions with others. It has been established that ethics are not a social construct [7], and that humans are born with some sense of morality.

2.2.5 Can we code it?

Is it possible to build an ethical AI? According to some AI experts, yes, we can, but in the future [10]. It is unknown when exactly we will have the capability to create ethical AI, but what we do know is that many AI applications that we have today, like autonomous cars, require us to code ethics into them, as they make moral decisions daily.

One obstacle we have to overcome when teaching morality to computers, is that we cannot objectively express morality in measurable parameters that they can process. Machines need explicit metrics, that can be measured and optimized. Another obstacle is how do we teach a machine to overcome biases, in terms of race, gender, sexuality etcetera? A machine knows what fair is based on what the programmers think fairness is [10], but we will explore this topic further in a future section.

So, after knowing some of the obstacles that we have to overcome, how do we approach this problem? Three ways have been proposed by machine learning experts [10].

2.2.5.1 Defining ethical behavior explicitly

Ethical values must be expressed by ethicists and AI researchers as quantifiable parameters. They have to provide answers for every ethical dilemma a machine might encounter, which would mean that they have to agree on what solution is most ethical for every problem. An example of this would be to code ethical values in automated cars, so they prioritize human life above all else. [10]

2.2.5.2 Crowdsourcing human morality

In order to properly train AI, we need to collect enough data on explicit ethical measures. No one situation is like the other, so we may need different ethical approaches. A solution to this is to crowdsource potential solutions form numerous people [10]. An excellent example of this is MIT's Moral Machine project, which is an online survey that collects data on how people would want autonomous vehicles to solve several moral dilemmas in the context of unavoidable accidents [11]. Another way

to see this, is trying to find a pattern in the choices we make and create AI that can predict what kind of decision a person would make in that situation.

2.2.5.3 Transparency at the heart of AI

When it comes to ethical matters, AI decisions should be more transparent in regard to ethical metrics and outcomes. We need more transparency about how ethics were quantified before being coded, and the outcomes AI produced as a result of those decisions. For example, to ensure ethical accountability, self-driving cars should keep detailed logs of all decisions they made. [10]

2.2.5.4 Problems with these suggestions

Unfortunately, these methods are not something that is set in stone but rather just some recommendations experts have made. Many problems arise, with the most important being that morality is not universal nor timeless, meaning that if we presented the same scenarios to different people in time and locations, we would probably get results that are racist, sexist and out right wrong considering today's standards. Apart from that though, if we take into consideration the different ethical theories, we saw in sections 2.2.1 – 2.2.3, we can see that we ran in to even more problems. The opinions of a strict Kantian deontologist and that of a strict utilitarian might be massively different, but still considered "correct" to themselves [13]. At most the outcome will be that they just have a difference of opinion and not that they are immoral or unethical. What happens though when the AI we create has a different opinion than our own? Do we deem it as "incorrect" and unethical, or do we just simply say that we have a difference of opinion?

Nevertheless, we are ignoring the elephant in the room. That if we want to create ethical AI, we have to find a way to quantify our values and explicitly define ethical rules, including the value of life against everything else. This has proved to be something extremely hard, but maybe this doesn't mean that it is impossible [13]. We discussed in a previous section that humans evolve their ethical judgement as life goes on, perhaps we could do the same with AI, as Ariela Tubert suggests [14]. Provide it with a starting

set of rules and values and through our feedback, lead it to the right direction. This would also mean that the AI will make a lot of mistakes along the way, just like humans do, before realizing that something is morally wrong. On the contrary, just like many kids learn that “lying is bad” but as they grow up, they realize that sometimes it is essential, the same can happen with AI [13].

2.3 Conclusion

We understand that AI systems make choices all the time and we want them to make the right ones. If we do not act now, we are making the moral choice of letting existing AI, make decisions that could probably be immoral and affect the lives of many people.

Though not perfect, we can use the suggestions that other experts made, to try and create ethical AI. This is not an easy ordeal as scientists are basically tasked with creating a “Good Samaritan AI” [10].

It is common understanding that humans are not born perfect, but it is also true that most of us try to do good every day. I believe we try to be good because of our bonds with other people, even if we disappoint them sometimes. If this is enough for us, why shouldn’t it be enough for AI? It will be designed by us, so it should stand to reason that it will also act like us, make mistakes, and learn from them. Expecting AI to be flawless is utopic, without having flawless humans to design it.

Chapter 3

Autonomous Vehicles

3.1 Definition

When hearing the term ‘autonomous vehicles’ we automatically think of a car with no driver. But this term can be broken down further than that. According to European guidelines, we can distinguish between six different levels of automated driving [1]:

0. No automations: this is the level where most cars are in our everyday life.
1. Driver assistance: this could be something simple, like brake assistance.
2. Partial driving automation: this includes braking, accelerating, and changing lanes. The driver has to be on alert and be prepared to take control whenever it is necessary.
3. Conditional driving automation: under certain circumstances the system can work autonomously for a certain amount of time before control is handed back to the driver.
4. High driving automation: the vehicle can perform all driving functions under standard circumstances and the driver is not required to take control.

5. Full driving automation: the vehicle can perform all driving functions in all circumstances. This means that the driver is no longer needed, and the car is fully autonomous.

As we can see, there is a noticeable change between level 2 and 3. The difference being that the driver is not required all the time. We will see later that this creates many problems but, in the meantime, we have to acknowledge that automated vehicles are the future. They will allow people who are not able bodied or people who are not that good at driving, to drive effortlessly and they will for sure revolutionize our transport system, which is currently not ideal.

It is important to note that most commercial cars that are available to the public are between the levels 0 and 2, such as the cars from Tesla [1]. Examples of upper levels include Waymo's and Cruise's driverless taxis in the US, Honda's and Mercedes-Benz level 3 car, and Toyota's level 4 service at the Tokyo 2020 Olympic Village [15].

Most autonomous vehicles use sensors, cameras and lasers which are located all around them, for detecting obstacles and use GPS to help them navigate [4]. All the aforementioned features give the vehicle the ability to have an accurate picture of its surroundings, and essentially allowing it to see if the car in front of it is suddenly slowing down or if a pedestrian suddenly jumps in the road. For this reason, these vehicles are expected to be much safer compared to the vehicles we drive today [16].

But how safe is "safe enough"? A question that is asked in the book Markus D. Dubber et al [16]. Autonomous vehicles will be programmed to drive "by the rules" [19], but even Google has acknowledged that they have allowed their vehicles to exceed the speed limit sometimes, as going slower would be more dangerous [25]. It has been suggested that an autonomous vehicle must drive 440 million kilometers without having an accident to be considered safe [1]. Even if this is true, and autonomous vehicles reduce traffic fatalities by a substantial amount, there would still be several lawsuits for the accidents they were not able to stop [5]. And who would be at the other end of the lawsuit and held responsible?

For so long, the driver has been the one responsible for their car. Logic would say that the accountability would shift from the driver to the manufactures and operators, since the driver's input is no longer required. This is where product liability comes into play, which is where a company has a legal responsibility to compensate for any damage their product has caused [1]. For this reason, researchers believe that a kind of 'black box' should be installed. Its purpose will be to record who is in control of the car at any given moment [1, 24], making it easier to shift liability between the driver and the manufacturers/operators.

But of course, the subject of autonomous vehicles comes with its own discourse. Questions like 'should autonomous vehicles be programmed to save their passengers at all costs?' led many philosophers to finding many similarities between this problem and the famous 'trolley problem'.

3.2 Trolley Problem

The 'trolley problem' is a thought experiment that involves an ethical dilemma where you are asked whether to sacrifice one person or save a larger number of people [17]. It has many variations, but the main scenario goes as such: there is a driverless trolley that is about to hit and kill a group of people but if you switch the tracks, the trolley will kill just one person that is on a different track. This can be seen in Figure 3.1. Another famous variation is one where you are at bridge above the tracks. In this case you have the choice to let the group of people die or throw over a large person that would be able to stop the trolley. This can be seen in Figure 3.2. As I said there are many variations including some that make the one person you have to sacrifice be someone you know (e.g., family, friends) and other variations that are more medical.

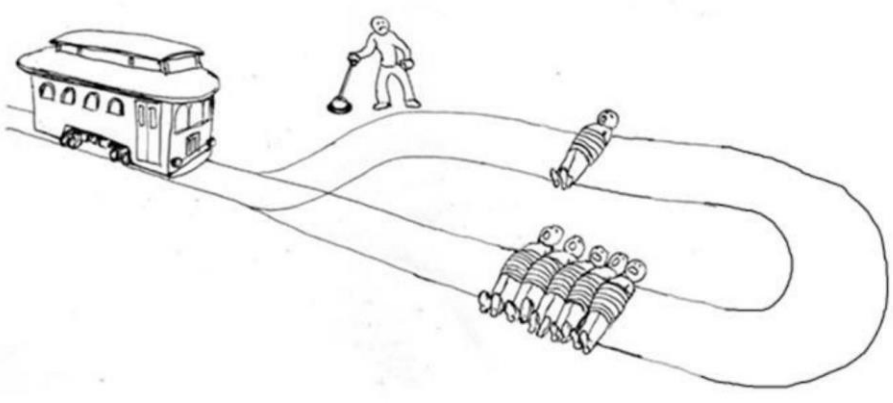

Figure 3.1: The trolley problem [49].

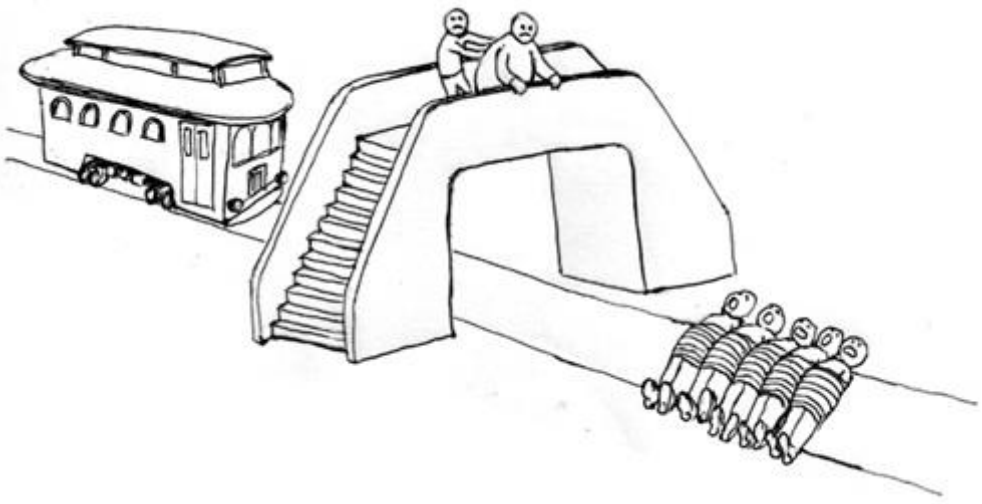

Figure 3.2: A variation of the trolley problem [50].

One of the most common responses for the first scenario is to switch the lever and kill the one person rather than the group of people, but when it comes to throwing the person over the bridge, people prefer to let the group die [18]. It is interesting that in the first case most are okay with sacrificing the one person but in the second they prefer to save the group. Why is it okay to sacrifice the one person using one method but not okay when using another? Let us flip the problem whereas a doctor can save a group of sick people who need organ transplants by using the organs of a healthy person. This does not include a trolley, but still falls under this category of thought experiments [18].

Philosophers use these kinds of thought experiments to investigate moral judgement and normative issues. Positive duties vs negative duties, killing vs letting die and consequentialism vs other approaches [18, 19]. It is important to note that these thought experiments are just that, experiments where choice is limited and where we have perfect knowledge of every outcome. They are not meant to be solved with a "right" choice [19].

3.3 Moral Machine

The 'Moral Machine' is game-like online platform developed by MIT with the goal of collecting data from the choices users make when deciding between two fatal outcomes of a moral dilemma. It has been compared to the aforementioned 'trolley problem' as the users have to decide whether the self-driving car continues its path or changes course. Either way people will die since the crash is inevitable. Scientists want to use the data that has been collected as research for decision algorithms. The platform can be accessed here. [20, 21]

It was mentioned in section 2.2.5.2, that a way we can program ethics into a machine is through crowdsourcing morality, which is what this online game tries to achieve. When the game starts you are presented with thirteen scenarios. For each one, you have to choose between two outcomes, where a description of each outcome is also given. This can be seen in Figure 3.3. But unlike the trolley problem where the only variables are the number of people killed and whether you had to do something with it or just let it happen, here there are many variables. Some of them include: the gender, age, physical condition (fit or unfit) and "social value" of passengers/pedestrians. What makes the game even harder is the inclusion of animals, pregnant people, and jaywalkers.

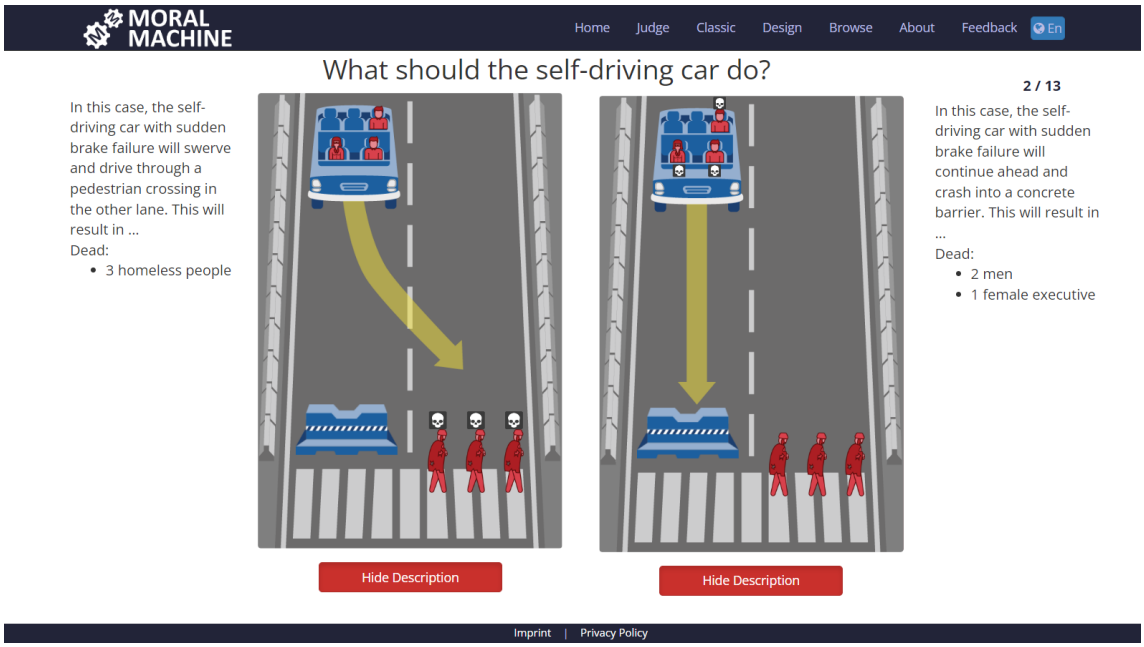

Figure 3.3: The Moral Machine [51].

After answering all thirteen dilemmas, you are presented with a screen where you can see which characters you saved the most and which you killed the most. You are also presented with some graphs which show how you fared compared to other people that have played this game. For example, you are shown how inclined you were to saving more lives, and also how much other people were. You are also given the chance to fill out a survey where you give more personal data like age, gender, highest level of education, political and religious views. You are also given different factors, like "saving more lives", and you have to answer how important each factor was to your judgment. In the end there are also some one-off questions like "how willing are you to buy a self-driving car" that you can answer.

Millions of people around the world have participated, and according to results that have been published most people prefer saving human lives, as many people as possible, children, fit and wealthy people. This means that they are more inclined to sacrificing old, overweight, or homeless people [21, 11].

The results also suggest that cultural traits and different moral principles are main reasons for the variety of ethical attitudes. Researchers want to use these results to design autonomous vehicles that correspond to the expectations and culture of where

they will operate, which further solidifies how generalizable ethical principles, and a universal ethical code are just mere pipedreams. [16]

This means that the solutions that are given by different ethicists for these kinds of dilemmas, might be rejected by the general public [16]. So, should ethicists ignore the views of the public and continue their tries on universal solutions for dilemmas? The answer is not that simple as both paths would be considered 'correct'. Majority preferences are not the most reliable indicators on what is correct and should be done [21]. A perfect example of this would be laws that protect minorities despite the general public being against this.

3.4 German Act on Autonomous Driving

The German Act on Autonomous Driving is a code of ethics that contains twenty ethical guidelines on how autonomous vehicles should act and should be programmed. The committee that created these guidelines consisted of professors of law, ethics, technical disciplines, and representatives of automotive companies. This was Germany's attempt to define ethical behavior explicitly and was also the first attempt worldwide to provide ethical guidelines for autonomous vehicles. [22]

The guidelines provide insight on subjects such as accountability, liability, data protection and many more. But I believe it is important to take a deep dive on the guidelines that refer to unavoidable dilemma decisions and general safety.

Starting with 'Ethical Guideline 2', which states:
"*The protection of individuals takes precedence over all other utilitarian considerations. The objective is to reduce the level of harm until it is completely prevented. The licensing of automated systems is not justifiable unless it promises to produce at least a diminution in harm compared with human driving, in other words a positive balance of risks*." [22]

This goes back to what was previously said about the safety of autonomous vehicles. They will and should be safer than conventional driving, or else there is no point in even

having autonomous vehicles. This guideline also mentions that the protection of individuals is far more important than any other advantage autonomous vehicles will bring along, such as minimizing traffic.

As mentioned in sections 3.2 and 3.3 a lot of the conversation that surrounds autonomous vehicles is about unavoidable crashes and dilemma situations, which the committee did not shy away from. The following guidelines address this topic directly.

'Ethical Guideline 5', which states:
"*Automated and connected technology should prevent accidents wherever this is practically possible. Based on the state of the art, the technology must be designed in such a way that critical situations do not arise in the first place. These include dilemma situations, in other words a situation in which an automated vehicle has to "decide" which of two evils, between which there can be no trade-off, it necessarily has to perform. ...*" [22]

This might be the second most important guideline. Instead of the countless hours spent arguing over the 'trolley problem', this guideline redirects the conversation in the direction it should be. Creating technology that ensures these 'dilemmas' will not occur in the first place.

'Ethical Guideline 7', which states:
"*In hazardous situations that prove to be unavoidable, despite all technological precautions being taken, the protection of human life enjoys top priority in a balancing of legally protected interests. Thus, within the constraints of what is technologically feasible, the systems must be programmed to accept damage to animals or property in a conflict if this means that personal injury can be prevented.*" [22]

'Ethical Guideline 8', which states:
"*Genuine dilemmatic decisions, such as a decision between one human life and another, depend on the actual specific situation, incorporating "unpredictable" behavior by parties affected. They can thus not be clearly standardized, nor can they be programmed such that they are ethically unquestionable. Technological systems must*

be designed to avoid accidents. However, they cannot be standardized to a complex or intuitive assessment of the impacts of an accident in such a way that they can replace or anticipate the decision of a responsible driver with the moral capacity to make correct judgements. It is true that a human driver would be acting unlawfully if he killed a person in an emergency to save the lives of one or more other persons, but he would not necessarily be acting culpably. Such legal judgements, made in retrospect and taking special circumstances into account, cannot readily be transformed into abstract/general ex ante appraisals and thus also not into corresponding programming activities. ..." [22]

'Ethical Guideline 9', which states:
"*In the event of unavoidable accident situations, any distinction based on personal features (age, gender, physical, or mental constitution) is strictly prohibited. It is also prohibited to offset victims against one another. General programming to reduce the number of personal injuries may be justifiable. Those parties involved in the generation of mobility risks must not sacrifice non-involved parties.*" [22]

These three guidelines might be considered the most important ones as they answer critical questions that were posed in previous sections. Protection of the human life is the priority and personal features should not play a role, which is contrary to what the 'Moral Machine' tried to do. It is also highlighted that the decision to sacrifice specific lives on purpose should not be taken by the programmers. Lastly, not permitting non-involved parties to be sacrificed indicates that the software cannot save the driver unconditionally.

Currently autonomous vehicles do not engage in trolley problem dilemmas. They are programmed to avoid any obstacles though. Cars already know how many passengers are being transported and wireless communication between cars is also available. So, in the event of a crash between two cars it would be possible for them to negotiate and give priority to the car with the most passengers. But this kind of offsetting is prohibited by Ethical Guideline 9. [1]

This act was revolutionary for the autonomous vehicle conversation and paves the way on how to legislate higher levels of autonomous driving. A big takeaway from this, is something that was mentioned previously, which is that we need to integrate ethics into autonomous vehicles and do it from the development stage. [23]

3.5 Where the Moral Machine goes wrong

Engineers are pointing out that the way this experiment frames the problem is wrong as it is just an extremely rare case [21, 19, 1]. It not only exaggerates the likelihood of such events, but it is also asking the wrong question completely. When presented with two potential outcomes of a scenario, humans will grasp onto whatever features are given, so they can find something that differentiates the two outcomes and make their choice easier. And most of the time the features that are offered, try to capture 'social value' [21].

Yet 'social value' is not something you can calculate just by looking at someone. The most renowned scientist might dress simpler or might even look like a homeless person. This simple example showcases how the strategy of trying to favor those of 'higher social value' is not practical or logical. It also ensures that the choice the user makes will be morally arbitrary and based solely on predetermined biases [21]. The Moral Machine encourages users to voice their biases and promotes the moral relevance of those biases, which leads to bad results. We should move away from the single instances of 'to swerve or not' and think about what the world would look like if we were to implement our preference, as Abby Everett Jaques suggests [21].

Let us imagine someone chooses to swerve to avoid hitting a kid and sacrifices an adult. What would it mean if we implemented this preference? Adults would always be afraid to be outside and would lose access to a lot of public places. This sounds preposterous and idiotic. Making a personal feature a key part of making a choice is a bad idea [21]. Though this way of thinking will not give us the actual solution, it at least moves the conversation forward and makes you think.

3.6 Why the trolley problem is not the best approach

We cannot deny that the ethical dilemma that is posed by autonomous vehicles is not similar to the trolley problem. At least on a surface level. But when diving deeper, we can easily understand they are two vastly different problems.

Let us start with the most obvious one. In the trolley problem thought experiment, both moral and legal responsibility are set aside and not taken into account. But when it comes to autonomous vehicles, both need to be taken into consideration. We also need to keep in mind that the decision making in the trolley problem is made by one person in the now, whereas autonomous vehicles will have pre-programmed decision making that was decided by a group of people.

In addition, it is crucial to point out that all the variations of the trolley problem are far removed from our reality. For example, nobody can be sure that pushing a large person in front of a trolley would stop it for sure in real life [18]. This is to point out that in the thought experiments there is perfect knowledge of everything, meaning that the outcome of every choice and the number of choices we have is known. But that does not translate to the real world. An autonomous vehicle cannot have perfect knowledge of the state of the road, just estimates [18]. This means that autonomous vehicles will make decisions under uncertainty, since not every outcome is totally certain and the pool of number of choices they have to make is not limited.

3.7 Conclusion

We understand that autonomous vehicles are the future of transportation. Unfortunately, they come with their own set of ethical concerns and implications. I personally would use autonomous vehicles as I believe they would make my life way more comfortable and easier. Before acquiring one though, I would like to know how the ethics behind it were programmed, meaning was it programed to always put the driver’s life as a priority, or programmed to save as many lives as possible even if that means the driver might not survive? I also believe that before acquiring one of these vehicles, the driver should be informed as to what happens in case of an accident and who will be held accountable. It is important to educate the drivers and general public on the pros and

cons of autonomous vehicles without fear mongering, which is how this subject has been handled up until now.

The public discourse and fear mongering on this subject has derailed the conversation towards the trolley problem. I too have fallen victim of this agenda and believed that by solving the trolley problem, we would also solve the ethical problems that come with autonomous vehicles. While the trolley thought experiment is a great conversation starter, it is far from the solution to this problem. It is a thought experiment that does not really have a right answer, as the 'right' answer depends on what ethical theory you subscribe to. The obsession with the 'trolley problem' has in turn created even more discourse with the introduction of the 'Moral Machine'.

In the planning phase of my thesis, I was considering creating a questionnaire that would present different trolley problem scenarios and ask people who they would rather sacrifice, as I believed this was a relevant and good way to find an answer to this ethical dilemma. After researching the topic further, I stumbled upon the 'Moral Machine' experiment. This was what I wanted to do, but rather on a global scale with more valid results. After playing this game and choosing who the autonomous vehicle should sacrifice, I did not feel moral or that my answers had helped provide a substantial solution to the dilemma. That is when I decided that maybe this approach is not the best, and to my surprise many experts had the same opinion. The only thing I achieved through my answers was to expose my biases, which surprised me as well since I believed I was a very unbiased person. The characteristics that helped differentiate the two dilemmas in each round, should not be the deciding factor as to who should be sacrificed, as having a preference to sacrifice older people rather than younger ones does not make you moral, but ageist. The 'Moral Machine' was a good effort to spark a conversation, but I am not sure how helpful it is for the future of autonomous vehicles.

Though the way the 'Moral Machine' was developed and executed is questionable to say the least, the process of crowdsourcing morality is not. I believe having the general public's view play a part on how accident algorithms are developed and what morals are imbued into them is a really promising idea, if done correctly. After all morality is not

something universal nor timeless, so the morals that are coded in machines should be ‘updated’ and be relevant to the current status of the world.

Both the ‘Moral Machine’ and trolley problem are the wrong way of approaching this delicate subject. What comes close though and gives hope for the future is Germany’s attempt on explicitly defining ethical guidelines for autonomous vehicles. These guidelines have done more to push the conversation forward, rather than the trolley problem comparison and the ‘Moral Machine’ experiment. I do not believe this act ended the discourse on the subject, but I certainly think it is a step forward, as it does not promote discrimination and biases and puts human life at the top. It definitely proves that having ethicists and other experts define ethical behavior explicitly, can work wonders. I agree with all the guidelines I highlighted in section 3.4, and I would go as far as to say that they should be even more precise and not as general.

Overall, both approaches of crowdsourcing morality and having ethicists explicitly define ethical guidelines can work if done right, and I believe that the combination of these two would lead to even greater results.

Chapter 4

The case of Google Duplex

4.1 Definition

Virtual assistants are software agents that use natural language processing and produce speech to help users with various tasks. They can help with various daily tasks, such as playing music, setting alarms, and reminding you of important events. Siri, the Google Assistant, and Amazon's Alexa are some of the most famous examples of these assistants. [26, 27]

These assistants have become a staple in peoples' lives as they help people get things done more easily. But in 2018, Google revolutionized the game. During one of their famous Google I/O conferences they unveiled a new feature of the Google Assistant called Google Duplex.

To demonstrate its capabilities, Duplex made a call to a hair salon. The goal of the call was to book an appointment for the user, without the user speaking at all. Not only did

Duplex manage to book the appointment, but it did so by including human like “hmms” and “umms”, and natural pauses in the conversation, just like humans do. This is a striking difference compared to its other counterparts like Alexa and Siri.

Duplex is only being trained to book a table at a restaurant or a hair salon appointment. This means that Duplex is unable to carry conversations outside its domains [30], which is something that helped Google overcome challenges like understanding natural language and modeling natural behavior [28].

When the user wants to book something, they provide Duplex with the necessary information (e.g., time, date, number of people) and it makes a call to the business. Apart from using speech disfluencies, it is able to use different tones, everyday expressions and even understand different contexts [30]. If the booking was successful, Duplex lets the user know the booking details. But if not, Duplex will call a human operator to take over. It is also important to note that all the conversations that Duplex makes, are being recorded [30].

Duplex surprised everyone with its very convincing human voice and natural conversation abilities. Google managed to create AI that was capable of having human conversations and sounding ‘natural’ or even human [28]. But of course, this excitement did not last for long. These assistants have always been under fire when it comes to ethicality, with issues like privacy, being recorded and always listening on the background [33]. But Duplex’s issues delve even deeper than that.

4.2 Where the Duplex goes wrong

The existence of Duplex made AI ethicists question whether Google had crossed a line [27]. During the call, the hair salon receptionist believed she was talking to a real human at the other end of the phone call. Did Google purposefully deceive the receptionist? A lot of people certainly believe so, since there is no reason for the human like “umms” and pauses, other than to deceive [27]. And in addition to that, no disclaimer was given to the receptionist that she was not talking to a real human being.

Even a lot of media outlets criticized the existence of Duplex, by calling it “scary” and many questioned whether we have approached an AI “tipping point” [28].

The fact that Duplex was able to deceive someone during a conversation made a lot of people wonder whether this means that it has passed the famous ‘Turing Test’, that was mentioned in section 2.1. It is yet unclear whether it has passed said test, but ironically, even the idea of it passing it, has raised some ethical concerns.

To pass the test, a human has to have a conversation with a machine, without understanding that they are talking to machine. Which is what happened at the Google I/O demonstration. But ethicists have agreed that when AI converses with humans, it has the duty to let them know that it is not a human [27]. Up until this point this was a bit redundant, as it was obvious by the way the AI was conversing that it was not human, due to its limited language understanding. And then Duplex came into the picture.

4.3 Policies

Even before the introduction of Duplex a lot of policies existed preventing the very thing that Duplex is being criticized for. For starters, California introduced a bill in 2018 that would make it “unlawful for any person to use a bot to communicate …, with the intent to mislead the other person about its artificial identity for the purpose of knowingly deceiving the person about the content of the communication” [31]. In addition, the IEEE recommends that “the AI’s artifactual nature should always be made as transparent as possible” [29].

Since these policies were in place before the existence of Duplex, it is questionable why Google has not abided by them. Though Google has been called out for not following the aforementioned guidelines [16].

4.4 From a Kantian approach

Like it was mentioned before, Duplex seems to lie to disguise its artifactual nature, which appears to be an important part of why this technology works in the first place. We know form Kant that lying is bad, no matter the outcome. Hence, this technology would be deemed morally wrong.

Many are also concerned about how Duplex can be used fraudulently. An ill-intentioned person can use Duplex's human like voice to speak on behalf of others and therefore committing a real-life crime, identity theft. Duplex's existence also belittles the last authentic way to identify an unknown person, the voice. [29]

Apart from identity theft, people are concerned that Duplex might make it even easier for businesses to get spammed and scammed [30]. During their presentation Google did not provide information on how businesses would consent to these automated calls, which is especially concerning when taking into account that the calls are being recorded by Duplex.

We have become used to communicating with bots as long as we initiate the conversation. But Duplex enters uncharted waters by having the bot initiate a conversation with us. This strange behavior decreases the well-being of a person [29]. Finally, Duplex creates a new worry, one where people have to doubt whether a human is behind every single phone call [30].

4.5 From a utilitarian approach

On the other hand, a utilitarian might argue that in order to maximize the Duplex's utility, deception is necessary. This is backed up by IEEE guidelines which mention that "deception may be acceptable … when it is used for the benefit of the person being deceived" and that "for deception to be used under any circumstance, a logical and reasonable justification must be provided" [29].

The point that deception is necessary, comes from the fact that people are known to feel uncanny when interacting with anthropomorphic bots. This has been referred to as the 'uncanny valley', which is a concept that claims that when humans interact with

anthropomorphic artificial bots that do not perfectly resemble humans, they get a strange and uneasy feeling [32]. This feeling passes when we cannot tell that we are communicating with a bot, which is what happens with Duplex, since Google designed it so that it would "make the conversation experience comfortable" [29]. By ridding the callee of this eerie feeling, Duplex promotes the well-being of the callee.

Another interesting point is that Duplex promotes efficiency. Since the only function of Duplex is to make reservations, it does not engage in small talk, and it does not discuss irrelevant topics. In turn, this allows the callee to do their job more efficiently. Unfortunately, researchers have found an 'efficiency - transparency' tradeoff, meaning that humans cooperate better with bots but only when they do not know of their artificial nature [29]. When that becomes known, performance is affected.

4.6 Does it make a difference if we know it is a robot?

Apart from the 'uncanny valley' that was mentioned before, would it really make such a difference if we knew the nature of the caller? Announcing the artificial nature of the caller, seems like the simplest solution that would solve the biggest ethic problem that surrounds Goggle's Duplex. Yet this solution raises even more questions.

Would people take a call if at the other end is a robot? The answer is that they will most likely hang up [34]. This makes sense, when taking into account that spamming calls are increasing exponentially.

Would people trust a call from a computer? You do not have reassurance from a robot that an actual human will show up at the appointment that was booked. And what happens when nobody shows up at Duplex appointments? It will create bad reputation for Duplex and dim reservations made by it as illegitimate [34].

4.7 Google's response

After the debut of Goggle's Duplex, a representative from Google revealed that as a company they believed that Duplex should inform people of its artificial nature, even

though they did not mention this during their presentation [28]. Had they done that, we would have been having a whole different conversation as is evident by their later release ‘Call Screen’.

Some weeks after their initial demonstration, Google revealed new updates that were implemented to Duplex. The most important of these updates being that at the beginning of a conversation with Duplex, it will identify itself as “Google’s automated booking service” and will also let the callee know that the conversation will be recorded [30]. As previously said, by mentioning the true nature of Duplex, people reported getting the eerie feeling I described [29].

Google also commented on the criticism they received about Duplex’s speech disfluencies. They supported their decision on keeping this feature since without it a lot of people would end the call immediately, as it would sound very robotic and computerized [30].

They also made sure to give emphasis on how human support is also a vital part of this project. It was mentioned before that Google personnel would take over if the booking was not successful, but now they will also take over if the callee does not consent to being recorded [30]. The personnel will also add the establishment at a ‘Do Not Call List’ to prevent future Duplex calls from happening [28].

Finally, to combat spamming, Google made sure to mention that there will be a daily limit on how many calls an establishment can receive from a Duplex and how many calls a Duplex can make. [28]

4.8 Call Screen

Months after the infamous demonstration of Duplex, Google announced a new feature for it. The name of this feature is ‘Call Screen’, and it is a new way for everyday people to screen phone calls from unknown numbers. Just like with the initial functionality of Duplex, Call Screen allows users to converse with unknows callers and have human like conversations.

But unlike the initial version of Duplex the user can switch to a normal phone call if they desire to do so, and they can also choose what Duplex replies as they have oversight. The main difference is that here, Duplex identifies its artificial nature from the beginning. This was definitely a defining reason on why the public's reaction was characterized as 'useful' rather than 'scary', which was the case with the initial version [28]. This change also raises the ethicality of using AI this way [28].

4.9 Ethical Guidelines

After all the reactions from the public, researchers, and ethicists, one can easily conclude that there should be ethical guidelines on how to design ethical voice assistants. Through my research I have found several authors [30, 33] that propose their own guidelines, and will now present some that I believe are truly helpful:

1. Voice assistants should always begin the conversation with identifying their artificial nature. This was one of the main ethical criticisms for the initial version of Duplex and is backed up by California's 'bot bill' and IEEE. It should also state on behalf of whom the call is being made.
2. The voice assistant should also state whether the conversation is being recorded and the recipient should be able to decline.
3. The voice assistant should only be able to call establishments that have not opted-out from this kind of calls.
4. If it is deemed necessary, or requested by the callee, a human should be able to take over the call.

Though limited, these guidelines come as a direct response to the criticisms that products like Duplex have received and were explored in sections 4.2 – 4.6.

4.10 Conclusion

The impact that voice assistants have is undisputed. Google's Duplex sure has great potential in saving time for booking appointments, but the ethical implications it creates should not be swept under the rug.

When Google made this announcement in 2018, I was really excited for this new feature as I found it to be really convenient for me and would really cut down on the time I spend on trying to book a table or an appointment. I did not know any better. But when I searched for the reaction of others on the internet I stumbled upon a wave of backlash. Soon enough I realized that people were right to be mad as Google Duplex seemed to pose a lot of ethical concerns.

After learning about the ethical concerns and implications and how they can be combated, I believe it should be expected from future AI to announce its artificial nature. Transparency is one of the most important ethical requirements, especially in AI applications. This implies to letting people know they are being recorded.

Thankfully, Google has listened and made the much-needed alterations to its product. If anything, this has taught us that we need ethical guidelines in everything that has to do with AI. I do not believe that this is the last we will hear of this kind of voice assistants, and Google is definitely not slowing down on adding new updates to Duplex. And who knows, maybe it will not be long before a robot calls us.

Chapter 5

Intelligent User Interfaces

5.1 Definition

Intelligent User Interfaces (IUI) are the combination of user interfaces and artificial intelligence. One of the most famous examples of IUIs is Microsoft Office Assistant called ‘Clippy’ [35]. The main goal of such interfaces is to better understand the needs of the user, adapt to the user’s preferences and aid the communication between computer and user. They can achieve this since they use facets of cognitive science, psychology, and human computer interaction [36].

Intelligent User Interfaces can be found in numerous places. Websites like Google use this kind of technology to promote advertisements that the user will like depending on what they viewed and clicked on in the past, thus creating a better user experience. But IUIs can also be found in hardware like phones. One example of this, is when brightness auto adjusts based on how the user adjusted it in the past. Another famous example of this is Netflix’s recommendation system. [36]

IUIs have been deemed as the key for having a more personalized experience, as they simplify processes such as, finding information which reduces the user’s cognitive load. Overall, they give the user a better experience as they provide more personalized ways of interacting. On the other hand, IUIs come with a lot of privacy concerns. [36]

5.2 Ethical Considerations

Many of these systems work by creating a user model based on all the information that was collected. Since this is done in the background most users do not know it is happening. If the user does not consent to it, it creates a violation to the user's privacy. The same goes for whether these technologies record the user to collect even more information.

It has been proved before by Edward Snowden that our data might be monitored by the government without our consent. [36]

5.3 Ethical Guidelines

Intelligent User Interfaces are a big part of Human Computer Interaction. HCI is believed to be the key to making sure IUI and a lot of other technologies operate in an ethical way. A solid way to design ethical IUI and mitigate a lot of privacy issues, by combining legislation, social norms, industry standards and numerous certifications [37].

Researchers believe that by combining User Experience Design and Artificial Intelligence we will be able to create trustworthy, safe, and reliable systems. They also believe that this combination will be the answer to Explainable Artificial Intelligence. [38]

Engineers believe that the best way to move forward is by creating explainable user interfaces. This explainability, will be based on visualization which is proven to prevent unnecessary errors and confusion. This will allow users to understand everything that is happening and make systems more reliable. To enhance this sense of reliability, bias testing should be done to lessen racial, gender and other biases. [38]

Microsoft has proposed eighteen guidelines on Human-AI Interaction and two of them are very relevant to our discussion. Namely Guideline 5:
"*Match relevant social norms.*

Ensure the experience is delivered in a way that users would expect, given their social and cultural context.

Example:
[Voice Assistants, Product #1] '[The assistant] uses a semiformal voice to talk to you - spells out "okay" and asks further questions.' " [39]

And Guideline 6:
"*Mitigate social biases.*
Ensure the AI system's language and behaviors do not reinforce undesirable and unfair stereotypes and biases.

Example:
[Autocomplete, Product #2] 'The autocomplete feature clearly suggests both genders [him, her] without any bias while suggesting the text to complete.' " [39]

Microsoft believes that this topic should be an ongoing research with high interest. They recognize that these two guidelines are only the tip of the iceberg when it comes to addressing fairness and more general ethical topics. They suggest that moving forward ethical guidelines should be evaluated by a diverse group of people as this has shown to identify issues that would otherwise be invisible. [39]

5.4 Conclusion

With how widespread and highly used Intelligent User Interfaces are, we have to consider all the ethical implications they create. Microsoft and other researchers have tried to address some of these by also creating some guidelines for future systems. But a lot more research has to be done.

I believe Human-Computer Interaction is an important part of this conversation, as the design of interfaces has a principal role in the ethicality of many systems. One of them being autonomous vehicles where, as it was mentioned in chapter 3, it must be clear at all times who is in control through the interface of the vehicle. Also, users should

always consent to their data being used for adaptation and automation as is the case for Netflix for example. Lastly, I agree that future ethical guidelines that will be developed for IUIs, should be developed by a diverse team.

Chapter 6

Responsible AI

6.1 Definition

Let us imagine that an algorithm of an autonomous vehicle adjusted the space between the vehicle and pedestrians by analyzing settlements from previous crashes. Meaning if the amount was high, the vehicle would leave a lot more space between itself and a pedestrian. Though this might seem reasonable and efficient, the algorithm ultimately penalizes poor people, as it is shown that pedestrians who were hit and settled for less was due to the fact they were living in low-income neighborhoods [25]. This means that the risk of poor people being hit, would be higher compared to others. Even though the programmers had good intentions when designing this algorithm, it is evident that the outcome is not fair.

Algorithms being unfair and biased is not something new, as there have been numerous examples in the past. This happens because not all aspects of Responsible AI are kept in mind during the design process, and because engineers are concerned with what the software does, rather than how it does what it does. [40]

Responsible AI should not discriminate against anyone based on attributes over which the user has no control over. For example, it should not reject someone who is up for a job, has the required skills and knowledge, but resides in a poor neighborhood. It should also be fair, free of bias and transparent. We can break Responsible AI down to three subcategories, each as important as the other: Fair AI, Explainable AI, and Accountable AI. [40]

6.2 Impact on society

The need for Responsible AI comes from the impact AI has had on our society so far. The European Parliament's study [24] on this topic takes a deep dive on these issues. We will talk about three of them: privacy, bias, and trust.

6.2.1 Privacy

As we have seen in the chapter about Intelligent User Interfaces, AI can have a significant impact on a person's privacy. IUI's are capable of collecting data about a user, such as their interests, which raises many concerns about whether they are always 'listening' in the background. When one considers the fear of being hacked, recorded all the time and their data being stored for a long time, it is understandable why there is such a big fear behind AI.

Lastly as mentioned by GDPR Article 21 "The data subject shall have the right to object, on grounds relating to his or her particular situation, at any time to processing of personal data concerning him or her…". [48]

6.2.2 Bias

When doing a deep dive on the 'Moral Machine' we saw how biased people can be. It stands to reason that since we are biased, AI that we create will also be. This might be due to the values of the AI engineers or the data that is being used to train said AI is reflecting only certain demographic groups.

An example of biased AI is COMPAS, which is software that calculates the probability of criminals to break the law again. It turns out that this was very biased against black Americans as it was more inclined to falsely predicted that they would break the law, compared to their white counterparts. An example of this can be seen in Figure 6.1.

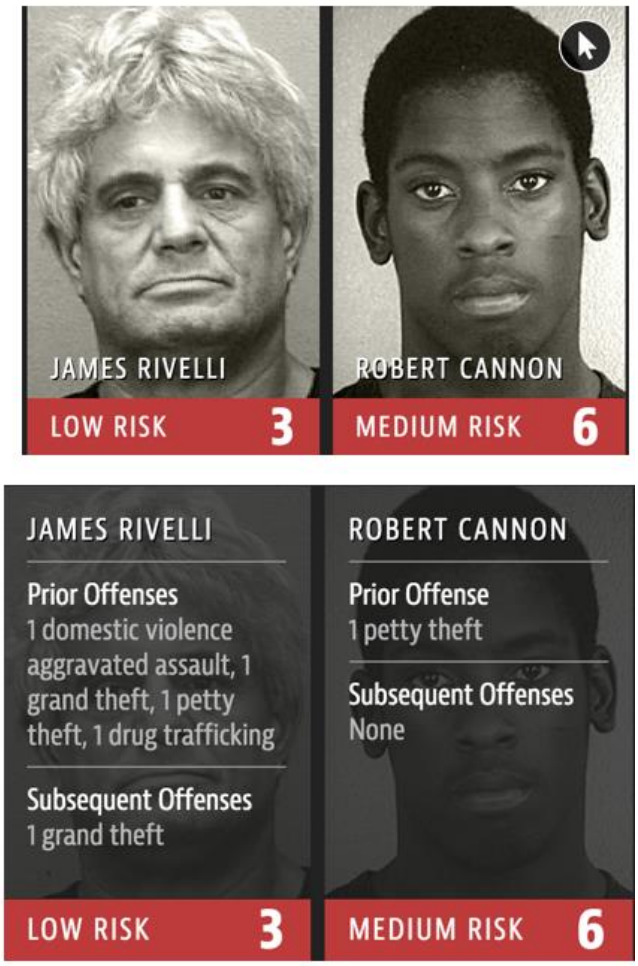

Figure 6.1: COMPAS example. [52]

More recently though in 2021, during one of their conferences, Google revealed that the cameras on their famous Pixel phones where biased towards people of color since they did not accurately portray them in photos. They go on to explain that since people of color were excluded during the testing period, the camera tends to overbright or desaturate their skin. An example of this can be seen in Figure 6.2. Google has reassured people that now their camera models will include more portraits of people of color. [41, 42]

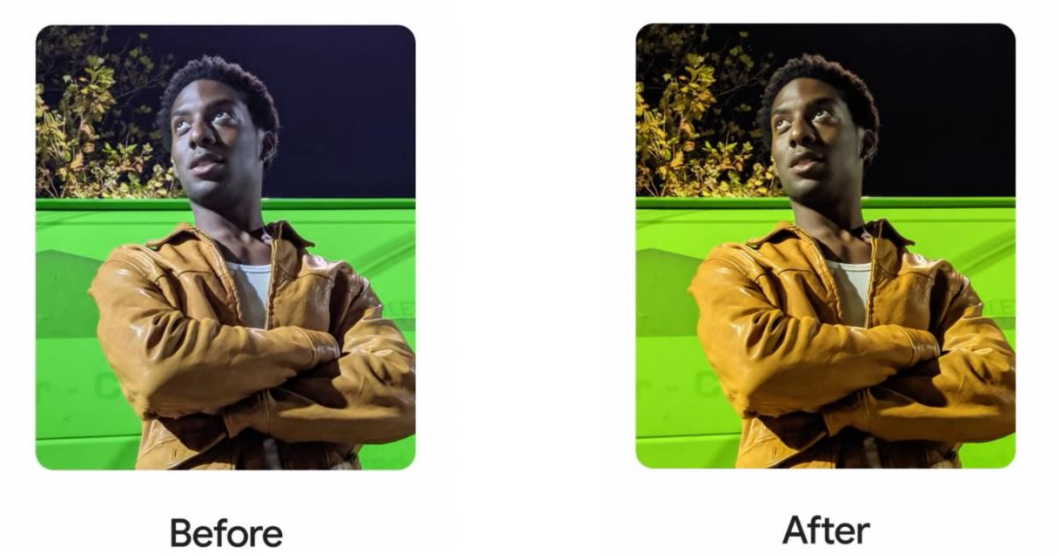

Figure 6.2: Google True Tone example. [53]

The most worrying part comes when we think about how AI can be used in law enforcement, which could result in certain groups of people being unfairly punished. Minorities could be marginalized even more considering the fact that AI is being used to screen people for university and job applications. These biases need to be eliminated in order for AI to be fair and responsible.

6.2.3 Trust

Trust is key for everything, and for AI to reach its full potential, we have to trust it. Constant supervision would render the use of AI useless. On the other hand, complete autonomy may lead to serious risks in safety.

Fairness is an essential part of trusting AI. Whatever the task, the decisions that AI makes should be fair and equal. We should also know why and how these decisions were made, which requires transparency and explainability, something that was explored during the discussion about Autonomous Vehicles and ‘how to program ethics’. The outcomes of these decisions should also be held accountable through regulations. We have to combine all these elements in order to achieve trustworthy AI.

6.3 Ethics Guidelines For Trustworthy AI

After understanding the impact that AI has in our lives the European Commission appointed an independent high-level expert group to create ethical guidelines [43] on how to develop trustworthy AI that is aligned with the European Union's foundational values. Though these guidelines refer to trustworthy AI, they also cover a lot of aspects of Responsible AI that were previously mentioned. These guidelines are very relevant to the discussion at hand and deal with a lot of the problems that were aforementioned throughout all the chapters.

The guidelines begin by defining that Trustworthy AI should be lawful, ethical, and robust. Each of these components is necessary to develop trustworthiness. The law provides a path on what should and should not be done and when that is not enough, we have ethical norms, and we also need to ensure that everything the AI does not cause any unintentional harm. It is important to note that the approach on defining these guidelines is 'human centric'.

The Commission uses five fundamental rights as the basis for Trustworthy AI. Namely these are:

1. Respect for human dignity
2. Freedom of the individual
3. Respect for democracy, justice, and the rule of law
4. Equality, non – discrimination and solidarity
5. Citizen's rights

By using these fundamental rights, they came up with four ethical principles that AI should always stick to. These are:

1. Respect for human autonomy
2. Prevention of harm
3. Fairness
4. Explicability

In turn, these principles were translated into tangible requirements to achieve Trustworthy AI. These requirements refer to developers, deployers and even end users and should be applied throughout the life cycle of an AI. The requirements are:

1. Human agency and oversight: Human autonomy and decision making should be promoted by AI and a central part of its functionality. This will be ensured by human oversight and governance mechanisms.
2. Technical robustness and safety: AI should be resilient to outside attacks and should also have a fallback plan in case something goes wrong. In addition, AI should be accurate on the decisions it makes and the results of said actions should be reliable and reproducible.
3. Privacy and data governance: AI should ensure privacy, data protection, and have pre-determined data protocols on who can access data.
4. Transparency: The process on how an AI decision was made should be documented and traceable. This will help with explainability, which states that the decisions that the AI made should be comprehensible by humans. Lastly AI should reveal its artificial nature when interacting with humans.
5. Diversity, non-discrimination, and fairness: All identifiable bias should be removed and hiring people of diverse background is encouraged. AI services should not follow a 'one size fits all' approach and should be usable by everyone no matter their gender, disabilities, etc.
6. Societal and environmental well-being: AI should achieve its goal in the most environmentally friendly way possible. It should also promote the well-being of humans and not deteriorate it.
7. Accountability: AI algorithms, data and design processes should be evaluated by external and internal auditors. The negative impacts of an AI system should be reported but also minimized. Compensation should be possible for when things might go wrong. Lastly while trying to implement some requirements, tradeoffs might occur. They should be evaluated based on how much risk they impose to the ethical principles that were defined previously.

The Commission also suggests some methods that will help with the implementation of these requirements. For starters they suggest a 'white-list' and 'black-list' of rules that would indicate rules the system should always follow and rules the system should never

disobey, respectively. Another method that I find quite interesting is to provide explicit links between the ethical principles and the various implementation decisions that were made. Lastly, they propose appointing a person to oversee ethical AI issues as well as hiring a more diverse and inclusive team.

Lastly, the Commission has proposed four different ways to categorize the risk of AI applications [47]. Firstly, 'Unacceptable' which encompasses any AI application that might be a threat to any European citizen such as social scoring by the government. The next category is 'High Risk' which includes transport systems, such as autonomous vehicles, AI applications in education, employment, and law enforcement. The second to last category is 'Limited Risk' where chatbots that have minimal transparency obligations can be found. Lastly there is 'Minimal Risk' which is the category where most AI applications fall into, such as spam filters. These different categories can be seen in the Figure 6.1.

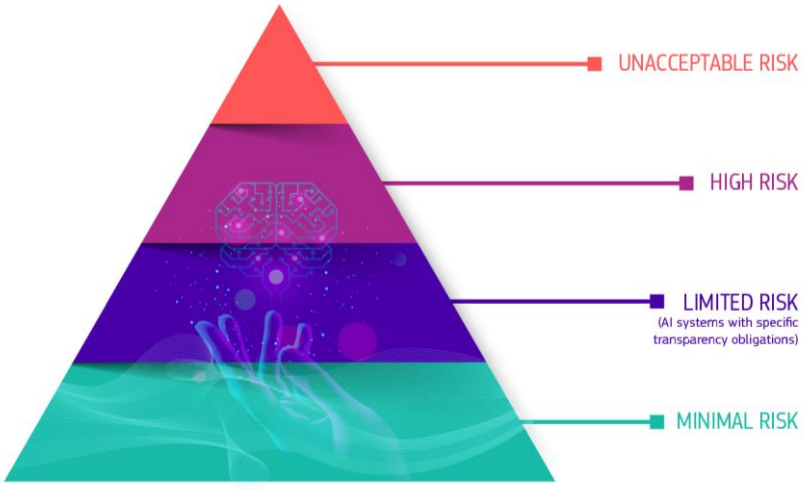

Figure 6.1: The pyramid that splits AI systems into four categories.
European Union, 2021 [47].

6.4 Education

The Commission also believes that education plays a big part in order to foster an ethical mind-set and that basic AI knowledge should be encouraged [43]. While on the topic of education, the Informatics for All coalition has proposed a plan to introduce

informatics in general education, as the skills and knowledge that will be gained from this subject are essential [44].

One of their main goals is for pupils of secondary education to be able to identify, analyze, and discuss the ethical and societal dilemmas that arise from the usage of computational systems, as well as their possible risks and benefits. Pupils will learn about the important ethical issues that come with the collection and use of users' data and how privacy plays a big role in that. They will also learn about some of the philosophical issues that surround AI and the ethical consequences that come with autonomous systems. Lastly pupils will explore numerous applications that have impacted our society in a big way, such as autonomous vehicles and intelligent personal assistants. This will give them the chance to highlight ethical issues but also understand different codes of ethics.

The use of AI in education is not a new concept as many universities use chatbots to support administrative and educational tasks [45]. A fellow student has developed a chatbot for his thesis project, where the goal of the chatbot is to help in education [46]. The chatbot is able to help students better understand the material they were taught, but also provide answers to questions that they have. In addition, it helps students with homework and assignments and can also provide further details about course fees, duration, and admission deadlines. The students at our university believe that this chatbot can, not only improve their education but also play a big part in it.

6.5 Conclusion

The evolution of AI during the last several years has brought forth the need to create not only Responsible but also Trustworthy AI. It is evident that the bias that programmers and developers have is passed onto AI algorithms. A way to ensure this bias in identifiable and not coded into algorithms in the first place is to have transparency throughout the development cycle of AI. This was also discussed on section 2.2.5.3, where it was suggested that transparency is the key for ethical AI. How bias is still a problem in 2022 is beyond me, but alas.

AI has impacted many facets of our society and daily lives. That is why it is essential to have guidelines like the ones that were proposed by the European Commission. I believe that these guidelines are the best attempt so far on trying to set rules for trustworthy AI and consequently ethical AI. I agree with all of them, and I will share some suggestions in the following chapter.

Lastly, I agree with the initiative of Informatics for All and believe that continuing on educating the public and especially young people on the ethical complications that AI brings, is a key factor in having ethical AI.

Chapter 7

Conclusion

7.1 Overview

Through this bibliographical thesis we have discussed several real-life AI applications and their respective ethical concerns, and how they impact our daily lives. We also explored numerous ways to combat these issues, some more successful than others.

After giving a definition for AI and giving a brief explanation on what ethics are and the three main ethical theories, we explored whether morality is something we are born with and if it something that someone can learn. We concluded that even though we are born with some sense of morality, we can also learn to be moral throughout our life, which in its turn indicates that computers can also learn. We then discussed three ways we can program ethics into machines. Namely by defining ethical behavior explicitly, crowdsourcing morality and having transparency be a key to the development of AI.

We continued with the first real life AI application that has many ethical concerns, Automated Vehicles. After explaining the different levels of autonomy, how autonomous cars work and addressing who is liable in case of an accident, we took a deep dive in the ‘trolley problem’ thought experiment and the ‘Moral Machine’. We saw how these two experiments were weaved into the conversation of autonomous vehicles and we explained why they do not provide any substantial guidance on how to solve the ethical issues that are created by autonomous vehicles. This turned our attention to Germany’s Act on Autonomous Driving which sparked an interesting conversation. Through the Moral Machine and Germany’s Act on Autonomous Driving

we saw two of the way we can code morality, crowdsourcing morality and defining ethical behavior explicitly. We concluded that though the Moral Machine does not do justice to crowdsourced morality, we can use as an example of what not to do in the future. We also saw how transparency can help car manufacturers know who was in charge of the car at all times through 'black boxes'.

We continued by discussing the ethical implications that came with the introduction of Google's Duplex and how the public and experts reacted. We also saw how Google responded to the backlash and how 'Screen Call' differentiates from Duplex. Lastly, we added some proposed guidelines to our conversation to make sure this does not happen again in the future. This showcased how important the third way of coding morality is, transparency at the heart of AI. We saw how the main ethical concern of Duplex was transparency in the sense that it did not let the user know of its artificial nature or that they were being recorded.

To add to the conversation about transparency, we defined what Intelligent User Interfaces are and explained how Human Computer Interaction can help us design more ethical AI. We saw how IUIs can have privacy concerns and how they can be addressed through transparency. Lastly, we saw how important IUIs and HCI is by explaining how they can be used to make sure that the driver of an autonomous vehicle always knows who is in control of the vehicle.

Lastly, we explained how big of a role bias plays into AI and how important transparency is in trying to solve this problem. We introduced Europe's Ethics Guidelines For Trustworthy AI and indicated that in conjunction with education, they are the most promising way of developing ethical AI.

7.2 Suggestions

After taking everything into consideration I believe that moving forward all three ways of coding ethics into machines should be used. Each of these ways is necessary but not sufficient in it of itself to achieve ethical AI. We saw two promising guidelines, Germany's Act on Autonomous Driving and Europe's Ethics Guidelines For

Trustworthy AI. While both are good attempts, I believe vital aspects are missing from them. Transparency on how the guidelines were formed is essential and something that we can see in Europe's attempt but not in Germany's. I also believe that the guidelines should also take into consideration the public's view on these moral matters. This can be done by crowdsourcing morality, as previously mentioned. We saw how the 'Moral Machine' failed, so researchers and engineers should come with a new and more effective way to crowdsourcing morality. To summarize I believe all three ways of coding ethics should be considered when coming up with next version of Ethical Guidelines. Lastly, I believe that HCI techniques should be used when designing these guidelines. We saw that Europe's Ethics Guidelines For Trustworthy AI were designed with a 'human centric' approach which is key in HCI. This led to a guideline on how AI systems should not be designed with a 'one-size for all approach' which is another key aspect of HCI.

Apart from the aforementioned, I strongly believe that AI education should be introduced to schools and that EPL341 should be reformulated to include the aspect of ethics along with AI. These will make sure that all future AI advances are criticized correctly by the general public and not fall into the trap of fear mongering. It will also make sure that the general public knows about privacy and transparency issues. Lastly, apart from including ethics in EPL341, I believe there should also be a dedicated lectures on bias, the importance of it and how it can be avoided. Though I am not an expert I believe these suggestions can help make better ethical AI.

Bibliography

[1] Bartneck, C., Lütge, C., Wagner, A. R., & Welsh, S. (2021). *An Introduction to Ethics in Robotics and AI*. Springer Publishing.

[2] *A DEFINITION OF AI: MAIN CAPABILITIES AND SCIENTIFIC DISCIPLINES*. (2018, December). European Commission.

[3] Dignum, V. (2020). *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way (Artificial Intelligence: Foundations, Theory, and Algorithms)* (1st ed. 2019 ed.). Springer.

[4] Tzafestas, S. G. (2016). *Roboethics: A Navigating Overview (Intelligent Systems, Control and Automation: Science and Engineering, 79)* (Softcover reprint of the original 1st ed. 2016 ed.). Springer.

[5] Boddington, P. (2017). *Towards a Code of Ethics for Artificial Intelligence (Artificial Intelligence: Foundations, Theory, and Algorithms)* (1st ed. 2017 ed.). Springer.

[6] Kizza, J. M. (2016). *Ethics in Computing: A Concise Module (Undergraduate Topics in Computer Science)* (1st ed. 2016 ed.). Springer.

[7] Rakić, V. (2021). *The Ultimate Enhancement of Morality*. Springer Publishing.

[8] Andre, C., & Velasquez, M. (1987). *Can Ethics be Taught?* Santa Carla University. https://www.scu.edu/mcae/publications/iie/v1n1/taught.html

[9] Beach, D. (2018, March 8). *Ethics, natural or learned behavior?* LinkedIn. https://www.linkedin.com/pulse/ethics-natural-learned-behavior-david-beach/

[10] Polonski, S., PhD. (2018, June 21). *Can we teach morality to machines? Three perspectives on ethics for artificial intelligence*. Medium. https://medium.com/@slavaxyz/can-we-teach-morality-to-machines-three-perspectives-on-ethics-for-artificial-intelligence-64fe479e25d3

[11] Awad, E., Dsouza, S., Kim, R. *et al.* The Moral Machine experiment. *Nature* **563,** 59–64 (2018). https://doi.org/10.1038/s41586-018-0637-6

[12] Creighton, J. (2016, July 1). *The Evolution of AI: Can Morality be Programmed?* Futurism. https://futurism.com/the-evolution-of-ai-can-morality-be-programmed

[13] Miller, K. (2017, September 14). *Can We Program Ethics into AI?* IEEE Technology and Society. https://technologyandsociety.org/can-we-program-ethics-into-ai/

[14] Ariela Tubert, *Ethical Machines?*, 41 SEATTLE U. L. REV. 1163 (2018).

[15] Wikipedia contributors. (2022, May 30). *Self-driving car*. Wikipedia. https://en.wikipedia.org/wiki/Self-driving_car

[16] Dubber, M., Pasquale, F., & Das, S. (2021). *Oxford Handbook of Ethics of AI*. Oxford University Press.

[17] Wikipedia contributors. (2022a, May 27). *Trolley problem*. Wikipedia. https://en.wikipedia.org/wiki/Trolley_problem

[18] Nyholm, S., & Smids, J. (2016). The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem? *Ethical Theory and Moral Practice*, *19*(5), 1275–1289. https://doi.org/10.1007/s10677-016-9745-2

[19] Müller, Vincent C., "Ethics of Artificial Intelligence and Robotics", *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/

[20] Wikipedia contributors. (2022a, May 8). *Moral Machine*. Wikipedia. https://en.wikipedia.org/wiki/Moral_Machine

[21] Jaques, A. E. (2019, March). *Why the moral machine is a monster*. https://robots.law.miami.edu/2019/wp-content/uploads/2019/03/MoralMachineMonster.pdf

[22] Luetge, C. The German Ethics Code for Automated and Connected Driving. *Philos. Technol.* **30,** 547–558 (2017). https://doi.org/10.1007/s13347-017-0284-0

[23] Kriebitz, A., Max, R. & Lütge, C. The German Act on Autonomous Driving: Why Ethics Still Matters. *Philos. Technol.* **35,** 29 (2022). https://doi.org/10.1007/s13347-022-00526-2

[24] *The ethics of artificial intelligence: Issues and initiatives*. (2020). European Parliament. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf

[25] N. J. Goodall, "Can you program ethics into a self-driving car?," in IEEE Spectrum, vol. 53, no. 6, pp. 28-58, June 2016, doi: 10.1109/MSPEC.2016.7473149.

[26] Wikipedia contributors. (2022a, April 19). *Virtual assistant*. Wikipedia. https://en.wikipedia.org/wiki/Virtual_assistant

[27] Bock, M. A. (2019, April 18). *Is Google Duplex too human? : exploring user perceptions of opaque conversational agents*. The University of Texas at Austin. https://repositories.lib.utexas.edu/handle/2152/74330

[28] Grevatt, N. (2018, December). *Google's Duplex and Deception through Power and Dignity*. https://aipavilion.github.io/docs/papers/duplex.pdf

[29] S. (2021, March 6). *Google Duplex: The Effects of Deception on Well-Being*. Dis/Connected. https://simonfischer.me/google-duplex-the-effects-of-deception-on-well-being/

[30] Pierantoni, F. (2021, December 8). *When AI speaks on behalf of humans: Proposing ethical guidelines based on Google Duplex assistant*. Medium. https://becominghuman.ai/when-ai-speaks-on-behalf-of-humans-proposing-ethical-guidelines-based-on-google-duplex-assistant-b9e5346a9aa5

[31] *Bill Text - SB-1001 Bots: disclosure.* (2018). California Legislative Information. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1001

[32] Wikipedia contributors. (2022d, May 28). *Uncanny valley*. Wikipedia. https://en.wikipedia.org/wiki/Uncanny_valley

[33] Chowdhury, S. K. (2019). *1 How do ethics influence Google Duplex's acceptance and the incoming wave of Proactive Voice Assistants?* https://www.researchgate.net/publication/336529668_How_do_ethics_influence_Google_Duplex's_acceptance_and_the_incoming_wave_of_Proactive_Voice_Assistants

[34] O'Leary DE. GOOGLE'S Duplex: Pretending to be human. Intell Sys Acc Fin Mgmt. 2019;26: 46–53. https://doi.org/10.1002/isaf.1443

[35] Wikipedia contributors. (2018, April 29). *Intelligent user interface*. Wikipedia. https://en.wikipedia.org/wiki/Intelligent_user_interface

[36] Amer, S. (2016). *Ethical concerns regarding the use of Intelligent User Interfaces*. http://worldcomp-proceedings.com/proc/p2016/ICM3954.pdf

[37] C. J. Hazard and M. P. Singh, "Privacy Risks in Intelligent User Interfaces," in IEEE Internet Computing, vol. 20, no. 6, pp. 57-61, Nov.-Dec. 2016, doi:10.1109/MIC.2016.116.

[38] Shneiderman, B. (2021, August 1). *Responsible AI: Bridging From Ethics to Practice*. August 2021 | Communications of the ACM. https://cacm.acm.org/magazines/2021/8/254306-responsible-ai/fulltext

[39] Microsoft. (2019). *Guidelines for Human-AI Interaction*. https://www.microsoft.com/en-us/research/uploads/prod/2019/01/Guidelines-for-Human-AI-Interaction-camera-ready.pdf

[40] Agarwal, S., & Mishra, S. (2021). *Responsible AI*. Springer Publishing.

[41] Sims, D. (2021, October 20). *Google introduces Real Tone to fight bias in camera tech*. TechSpot. https://www.techspot.com/news/91834-google-introduces-real-tone-fight-bias-camera-tech.html

[42] Google. (2021). *Real Tone on Google Pixel*. https://store.google.com/intl/en/discover/realtone/

[43] EUROPEAN COMMISSION. (2019). *ETHICS GUIDELINES FOR TRUSTWORTHY AI*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[44] Informatics For All. (2022). *Informatics Reference Framework for School*. https://www.informaticsforall.org/wp-content/uploads/2022/03/Informatics-Reference-Framework-for-School-release-February-2022.pdf

[45] Gutierrez y Restrepo, Emmanuelle & Baldassarre, Martín & G. Boticario, Jesus. (2019). ACCESSIBILITY, BIASES AND ETHICS IN CHATBOTS AND INTELLIGENT AGENTS FOR EDUCATION. 10.21125/edulearn.2019.2196.

[46] Christoforos Efstathiou (May 2022), DEVELOPING AN AI-BASED CHATBOT FOR DIGITAL EDUCATION

[47] *Excellence and trust in artificial intelligence*. (2019). European Commission - European Commission. https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/excellence-trust-artificial-intelligence_en

[48] *Art. 21 GDPR – Right to object*. (2018, March 28). General Data Protection Regulation (GDPR). https://gdpr-info.eu/art-21-gdpr/

[49] *Trolley problem*. (2017). [Illustration]. http://shikharsachdev.com/trolley-problem/

[50] *Trolley Problem variation*. (n.d.). [Illustration]. https://sites.google.com/site/has233aw/the-trolley-problem

[51] *Moral Machine*. (n.d.). [Illustration]. https://www.moralmachine.net/

[52] *COMPAS*. (2018). [Illustration]. https://medium.com/thoughts-and-reflections/racial-bias-and-gender-bias-examples-in-ai-systems-7211e4c166a1

[53] *Real Tone*. (2021). [Illustration]. https://www.makeuseof.com/real-tone-pixel-camera/